垃圾短信识别

背景描述及需求

随着通讯时代的到来，手机成为人们日常通讯必不可少的工具之一。手机短信也以其操作简单、方便快捷等诸多优点，逐渐成为用户间沟通的重要桥梁，但在为人们带来极大便利的同时，随之而来的诸多垃圾短信问题日益严峻，广告信息、欺诈短信、谣言散布等短信内容，严重危害了社会公共安全。据360手机卫士安全中心发布的数据，2019年全国垃圾短信拦截总量约为320亿条，平均每天收到垃圾短信超过5000万条。据统计，各类垃圾内容中，冒充类短信占所有垃圾短信的92.8%，多以银行诈骗、互联网虚假网购、电信诈骗等内容为主。垃圾短信泛滥，已经严重影响到人们正常生活、运营商形象乃至社会稳定。

面对激烈的市场竞争，各大运营商及相关部门都在寻求一种快速、有效的垃圾短信识别方法。通过垃圾短信的精准识别，以完善用户的通讯环境，为有关部门提供有效依据，维护运营商利益。试通过建立合理的短信识别模型，对垃圾短信进行识别，解决运营商和手机用户等困扰。

现状分析

目前我国垃圾短信的现状：

1）垃圾短信黑色利益链：由于短信的方便、低成本等特点，垃圾短信已经形成了黑色利益链，严重为害社会公众安全。由于监管缺失，一些不良组织通过各式各样的渠道收集个人手机信息，并将手机信息卖给有需求的商家和业务人员获取利益，同时商家等通过发送广告推销、诈骗等垃圾短息，来谋取利益，严重危害了短信用户的信息安全及正常生活。

2）缺乏法律保护：目前，虽然我国颁布了有关调整或者规范短信的法律、法规、政策性文件，如公安部、信息产业部、中国银行业监督管理委员会联合发出《在全国范围内统一严打手机违法短信息的通知》等，但是以规范短信业务的制度法来说，仍属空白。对垃圾短信的识别并追踪，找到垃圾短信发送的种子用户，不仅能够打击该类违法分子，还能弥补法律空白。

3）短信内容日益多变：垃圾短信的投放方式和内容的不断改进及变化，导致垃圾短信的拦截效果越来越差，传统的基于敏感词过滤方法不仅易受到同义词、字符等干扰，且不能很好适应垃圾短信的快速变化。

因此，垃圾短信过滤方法的改进优化成为运营商需要重点解决的问题。

思路流程

垃圾短信识别的主要步骤如下所示：

1.数据获取，获取所需数据集；

2.数据预处理，对数据进行文本中文分词、停用词过滤处理等；

3.建模准备，将分词结果分别转换成文档-词条矩阵，并划分测试集与训练集；

4.模型构建与评价，构建随机森林模型，并建立评价指标精确率、召回率、F1值对模型分类效果进行评价。

5.分析结果，总结和建议。

实施过程

数据接入

目前，某运营商已经积累了大量的垃圾短信数据。本案例收集了295755条短信文本数据，字段说明如下：

字段名称	类型	字段说明
_c1	整型	0表示正常短信，1表示垃圾短信
_c2	字符串	短信内容

经过加工处理数据如图所示：

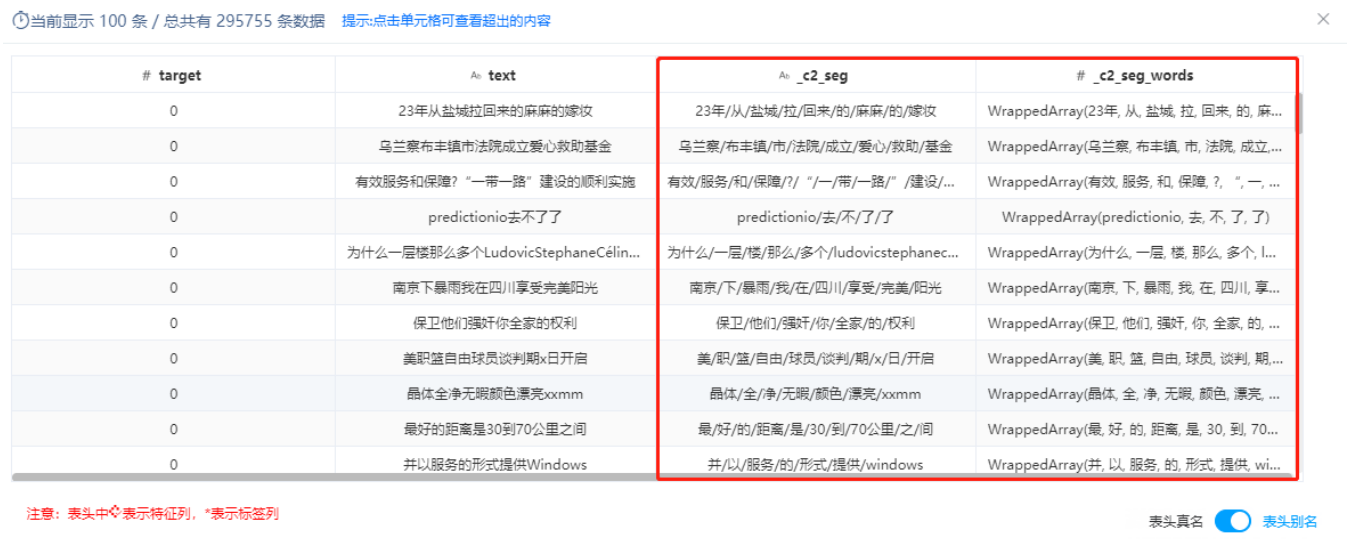
为了方便识别字段含义，这里接入一个 元数据编辑 节点取别名，如图：

数据预处理

1、分词

中文分词是指将一整段文字切分为具有最小语义的词条信息，即以词作为基本单元，使用计算机自动对中文文本进行词语的切分，将文本数据转化为机器可识别的形式。英文单词之间是由空格作为分界符的，中文则是由字为基本书写单位，词语之间没有明显的区分符，因此，中文分词是中文信息处理的基础与关键。分词结果的准确性，对后续文本挖掘有着重要影响。如在进行特征的选择时，不同的分词效果将影响词语在文本中的重要性，从而影响特征的选择。

这里接入一个 分词 节点将text列进行分词，_c2_seg为分词后的字符串型结果，_c2_seg_words为分词后的WrappedArray类型结果，分词输出结果如图：



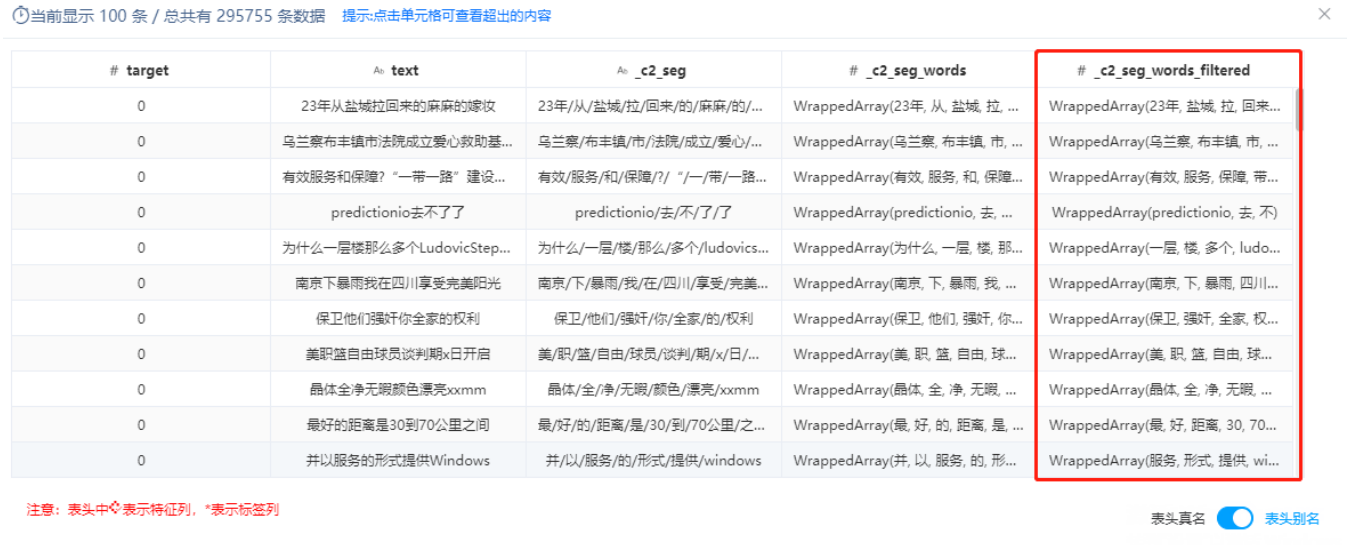
# target	text	_c2_seg	_c2_seg_words
0	23年从盐城拉回来的麻麻的嫁妆	23年/从/盐城/拉/回来/的/麻麻/的/嫁妆	WrappedArray(23年, 从, 盐城, 拉, 回来, 的, 麻...
0	乌兰察布丰镇市法院成立爱心救助基金	乌兰察/布丰镇/市/法院/成立/爱心/救助/基金	WrappedArray(乌兰察, 布丰镇, 市, 法院, 成立,...
0	有效服务和保障？“一带一路”建设的顺利实施	有效/服务/和/保障/?/“/一/带/一路/”/建设/...	WrappedArray(有效, 服务, 和, 保障, ?, “, 一, ...
0	predictionio去不了了	predictionio/去/不/了/了	WrappedArray(predictionio, 去, 不, 了, 了)
0	为什么一层楼那么多个LudovicStephaneCélin...	为什么/一层/楼/那么/多个/ludovicstephanec...	WrappedArray(为什么, 一层, 楼, 那么, 多个, l...
0	南京下暴雨我在四川享受完美阳光	南京/下/暴雨/我/在/四川/享受/完美/阳光	WrappedArray(南京, 下, 暴雨, 我, 在, 四川, 享...
0	保卫他们强奸你全家的权利	保卫/他们/强奸/你/全家/的/权利	WrappedArray(保卫, 他们, 强奸, 你, 全家, 的, ...
0	美职篮自由球员谈判期x日开启	美/职/篮/自由/球员/谈判/期/x/日/开启	WrappedArray(美, 职, 篮, 自由, 球员, 谈判, 期,...
0	晶体全净无暇颜色漂亮xxmm	晶体/全/净/无暇/颜色/漂亮/xxmm	WrappedArray(晶体, 全, 净, 无暇, 颜色, 漂亮, ...
0	最好的距离是30到70公里之间	最/好/的/距离/是/30/到/70公里/之/间	WrappedArray(最, 好, 的, 距离, 是, 30, 到, 70...
0	并以服务的形式提供Windows	并/以/服务/的/形式/提供/windows	WrappedArray(并, 以, 服务, 的, 形式, 提供, wi...

2、停用词处理

中文表达中常常包含许多功能性词语，相比于其它词汇，功能性词语并没有太多的实际含义。最常用的功能性词语是限定词，如“的”、“一个”、“这”、“那”等。这些词语的使用较大的作用仅仅是协助一些文本的名词描述和概念表达。在信息检索中，为节省存储空间和提高搜索效率，在处理自然语言数据(或文本)之前或之后会自动过滤掉某些字或词，这些字或词即被称为停用词。

我们选择_c2_seg_words列，接入一个 停用词处理 节点，自定义停用词列表如图：

输出结果如图：



当前显示 100 条 / 总共有 295755 条数据　提示:点击单元格可查看超出的内容

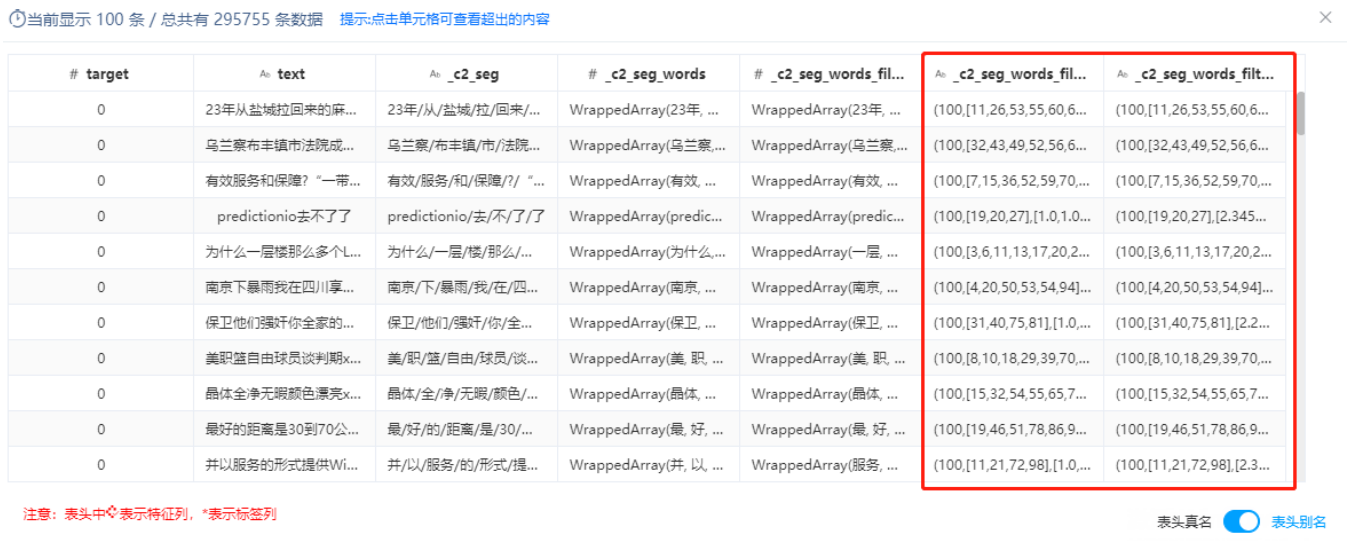
# target	text	_c2_seg	_c2_seg_words	_c2_seg_words_filtered
0	23年从盐城拉回来的麻麻的嫁妆	23年/从/盐城/拉/回来/的/麻麻/的/...	WrappedArray(23年, 从, 盐城, 拉, ...	WrappedArray(23年, 盐城, 拉, 回来...
0	乌兰察布丰镇市法院成立爱心救助基...	乌兰察/布丰镇/市/法院/成立/爱心/...	WrappedArray(乌兰察, 布丰镇, 市, ...	WrappedArray(乌兰察, 布丰镇, 市, ...
0	有效服务和保障？“一带一路”建设...	有效/服务/和/保障/?/ “/一/带/一路...	WrappedArray(有效, 服务, 和, 保障...	WrappedArray(有效, 服务, 保障, 带...
0	predictionio去不了了	predictionio/去/不/了/了	WrappedArray(predictionio, 去, ...	WrappedArray(predictionio, 去, 不)
0	为什么一层楼那么多个LudovicStep...	为什么/一层/楼/那么/多个/ludovics...	WrappedArray(为什么, 一层, 楼, 那...	WrappedArray(一层, 楼, 多个, ludo...
0	南京下暴雨我在四川享受完美阳光	南京/下/暴雨/我/在/四川/享受/完美...	WrappedArray(南京, 下, 暴雨, 我, ...	WrappedArray(南京, 下, 暴雨, 四川...
0	保卫他们强奸你全家的权利	保卫/他们/强奸/你/全家/的/权利	WrappedArray(保卫, 他们, 强奸, 你...	WrappedArray(保卫, 强奸, 全家, 权...
0	美职篮自由球员谈判期x日开启	美/职/篮/自由/球员/谈判/期/x/日/...	WrappedArray(美, 职, 篮, 自由, 球...	WrappedArray(美, 职, 篮, 自由, 球...
0	晶体全净无暇颜色漂亮xxmm	晶体/全/净/无暇/颜色/漂亮/xxmm	WrappedArray(晶体, 全, 净, 无暇, ...	WrappedArray(晶体, 全, 净, 无暇, ...
0	最好的距离是30到70公里之间	最/好/的/距离/是/30/到/70公里/之...	WrappedArray(最, 好, 的, 距离, 是, ...	WrappedArray(最, 好, 距离, 30, 70...
0	并以服务的形式提供Windows	并/以/服务/的/形式/提供/windows	WrappedArray(并, 以, 服务, 的, 形...	WrappedArray(服务, 形式, 提供, wi...

注意：表头中◇表示特征列，*表示标签列　　表头真名 表头别名

3、TF-IDF

由于文本数据无法直接用于建模，因此需要将文本表示成计算机能够直接处理的形式，即文本数字化。TF-IDF算法即将文本数据进行数值化。TF意思是词频，IDF意思是逆文本频率指数，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。TF-IDF值越高，说明该词越重要。

我们接入 TF-IDF 算法进行抽取变换，输出结果如图：



当前显示 100 条 / 总共有 295755 条数据　提示:点击单元格可查看超出的内容

# target	text	_c2_seg	_c2_seg_words	_c2_seg_words_fil...	_c2_seg_words_fil...	_c2_seg_words_filt...
0	23年从盐城拉回来的麻...	23年/从/盐城/拉/回来/...	WrappedArray(23年, ...	WrappedArray(23年, ...	(100,[11,26,53,55,60,6...	(100,[11,26,53,55,60,6...
0	乌兰察布丰镇市法院成...	乌兰察/布丰镇/市/法院...	WrappedArray(乌兰察,...	WrappedArray(乌兰察,...	(100,[32,43,49,52,56,6...	(100,[32,43,49,52,56,6...
0	有效服务和保障？“一带...	有效/服务/和/保障/?/ “...	WrappedArray(有效, ...	WrappedArray(有效, ...	(100,[7,15,36,52,59,70,...	(100,[7,15,36,52,59,70,...
0	predictionio去不了了	predictionio/去/不/了/了	WrappedArray(predic...	WrappedArray(predic...	(100,[19,20,27],[1.0,1.0...	(100,[19,20,27],[2.345...
0	为什么一层楼那么多个L...	为什么/一层/楼/那么/...	WrappedArray(为什么,...	WrappedArray(一层, ...	(100,[3,6,11,13,17,20,2...	(100,[3,6,11,13,17,20,2...
0	南京下暴雨我在四川享...	南京/下/暴雨/我/在/四...	WrappedArray(南京, ...	WrappedArray(南京, ...	(100,[4,20,50,53,54,94]...	(100,[4,20,50,53,54,94]...
0	保卫他们强奸你全家的...	保卫/他们/强奸/你/全...	WrappedArray(保卫, ...	WrappedArray(保卫, ...	(100,[31,40,75,81],[1.0,...	(100,[31,40,75,81],[2.2...
0	美职篮自由球员谈判期x...	美/职/篮/自由/球员/谈...	WrappedArray(美, 职, ...	WrappedArray(美, 职, ...	(100,[8,10,18,29,39,70,...	(100,[8,10,18,29,39,70,...
0	晶体全净无暇颜色漂亮x...	晶体/全/净/无暇/颜色/...	WrappedArray(晶体, ...	WrappedArray(晶体, ...	(100,[15,32,54,55,65,7...	(100,[15,32,54,55,65,7...
0	最好的距离是30到70公...	最/好/的/距离/是/30/...	WrappedArray(最, 好, ...	WrappedArray(最, 好, ...	(100,[19,46,51,78,86,9...	(100,[19,46,51,78,86,9...
0	并以服务的形式提供Wi...	并/以/服务/的/形式/提...	WrappedArray(并, 以, ...	WrappedArray(服务, ...	(100,[11,21,72,98],[1.0,...	(100,[11,21,72,98],[2.3...

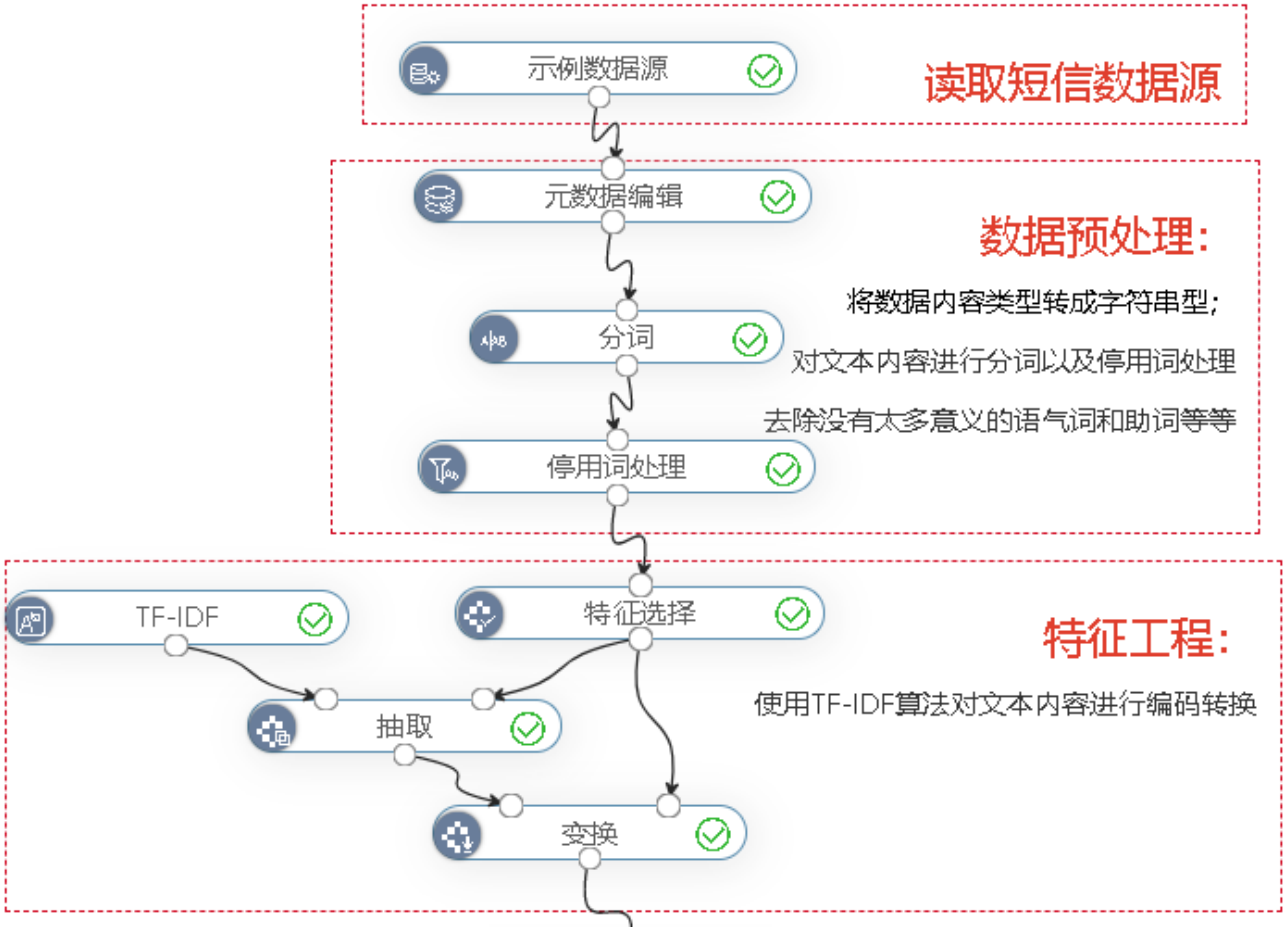
注意：表头中◇表示特征列，*表示标签列　　表头真名 表头别名

整个的数据预处理流程图如下：

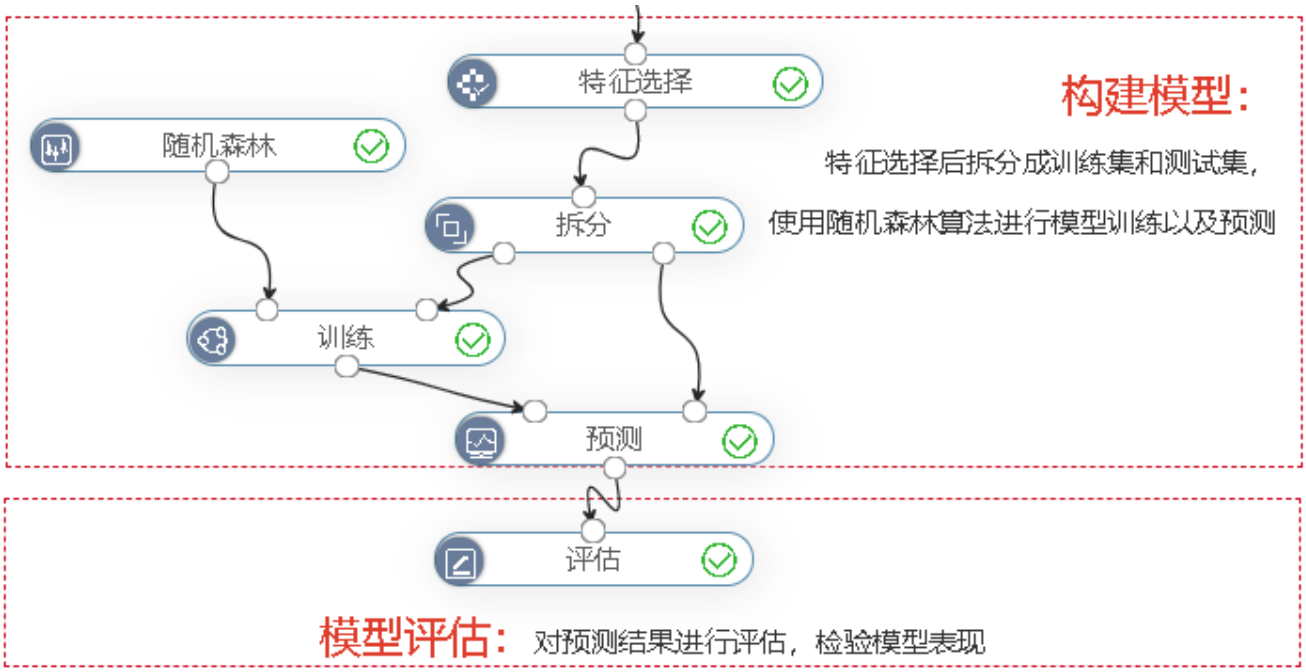
# target	text	_c2_seg	# _c2_seg_words	# _c2_seg_words_filter...	_c2_seg_words_filter...	_c2_seg_words_filter...
0	23年从盐城拉回来的麻麻...	23年/从/盐城/拉/回来/的/...	WrappedArray(23年, 从, ...	WrappedArray(23年, 盐...	(100,[4,32,40,47,53,57],[...	(100,[4,32,40,47,53,57],[...
0	乌兰察布丰镇市法院成立爱...	乌兰察/布丰镇/市/法院/成...	WrappedArray(乌兰察, 布...	WrappedArray(乌兰察, 布...	(100,[6,18,32,35,55,63,75...	(100,[6,18,32,35,55,63,75...
0	有效服务和保障？"一带一...	有效/服务/和/保障/?/ "/一...	WrappedArray(有效, 服务...	WrappedArray(有效, 服务...	(100,[5,37,41,57,71,77,90...	(100,[5,37,41,57,71,77,90...
0	predictionio去不了了	predictionio/去/不/了/了	WrappedArray(predictio...	WrappedArray(predictio...	(100,[27,28,59],[1.0,1.0,1....	(100,[27,28,59],[2.16020...
0	为什么一层楼那么多个Lud...	为什么/一层/楼/那么/多个/...	WrappedArray(为什么, 一...	WrappedArray(一层, 楼, ...	(100,[3,11,12,27,28,32,35...	(100,[3,11,12,27,28,32,35...
0	南京下暴雨我在四川享受完...	南京/下/暴雨/我/在/四川/...	WrappedArray(南京, 下, ...	WrappedArray(南京, 下, ...	(100,[27,39,58,67,89,94],...	(100,[27,39,58,67,89,94],...
0	保卫他们强奸你全家的权利	保卫/他们/强奸/你/全家/的...	WrappedArray(保卫, 他们...	WrappedArray(保卫, 强奸...	(100,[11,15,26,65],[1.0,1....	(100,[11,15,26,65],[2.003...
0	美职篮自由球员谈判期x日...	美/职/篮/自由/球员/谈判/...	WrappedArray(美, 职, 篮, ...	WrappedArray(美, 职, 篮, ...	(100,[3,4,12,22,25,56,86,...	(100,[3,4,12,22,25,56,86,...
0	晶体全净无瑕颜色漂亮xxmm	晶体/全/净/无瑕/颜色/漂亮...	WrappedArray(晶体, 全, ...	WrappedArray(晶体, 全, ...	(100,[20,39,60,69,76,80,8...	(100,[20,39,60,69,76,80,8...
0	最好的距离是30到70公里...	最好/的/距离/是/30/到/70...	WrappedArray(最好, 的, ...	WrappedArray(最好, 距离...	(100,[45,52,60,86],[1.0,1....	(100,[45,52,60,86],[1.825...
0	并以服务的形式提供Wind...	并/以/服务/的/形式/提供/...	WrappedArray(并, 以, 服...	WrappedArray(服务, 形式...	(100,[12,21,37,60],[1.0,1....	(100,[12,21,37,60],[2.093...
0	是大姑娘了要一点一点成熟...	是/大姑娘/了/要/一点一点/...	WrappedArray(是, 大姑娘...	WrappedArray(大姑娘, 一...	(100,[7,16,31,37,60,78,86...	(100,[7,16,31,37,60,78,86...
0	CancerNetwork总结了几...	cancernetwork/总结/了/...	WrappedArray(cancerne...	WrappedArray(cancerne...	(100,[2,3,8,20,69,80,92],[...	(100,[2,3,8,20,69,80,92],[...
0	先百度一下人工剥离胎盘是...	先/百度/一下/人工/剥离/胎...	WrappedArray(先, 百度, ...	WrappedArray(先, 百度, ...	(100,[5,9,18,40,49,74,91,...	(100,[5,9,18,40,49,74,91,...

构建模型

本案例采用 **随机森林** 算法模型，通过特征选择_c2_seg_words_filtered_idf列，目标标签为target，整体模型训练预测如图：



随机森林参数配置如图：



模型评价

通过 评估 节点接入，如构建模型的整体模型训练预测图所示。评估结果如图：

指标	值
confusion matrix(混淆矩阵)	真实|预测 0 1 0 63962 758 1 6565 17321
accuracy(准确率)	0.9173532266437939
roc曲线	查看ROC曲线
auc	0.9595720920639002
ks	查看KS曲线
weighted precision(加权精确率)	0.9207059731992495
weighted recall(加权召回率)	0.9173532266437938
weighted F1 score(加权F1分数)	0.9134092828766747
Class 0.0 precision(精确率)	0.9069150821671133
Class 0.0 recall(召回率)	0.9882880098887515
Class 0.0 F1 score(F1分数)	0.9458546215442857
Class 0.0 falsePositiveRate(假阳率)	0.2748471908230763
Class 0.0 truePositiveRate(真阳率)	0.9882880098887515
Class 1.0 precision(精确率)	0.9580729022622932
Class 1.0 recall(召回率)	0.7251528091769237

分析结果得出F1分数达到0.91,说明该模型效果比较不错的。

该模型能较好地识别出垃圾短信，有效进行垃圾短信过滤，解决运营商及用户的困扰。并且由上述分析提出以下建议：

- 对于垃圾短信过滤可结合传统匹配方法与基于内容的分类方法，不断优化识别模型以适应垃圾短信内容形式的不断变化。
- 对于垃圾短信泛滥问题，应当健全法律机制，结合垃圾短信识别系统，从根源上进行遏制，从而建立一个良好的通信环境。

总结

本案例运用短信数据，对垃圾短信进行识别。重点介绍了文本数据的处理及转换过程，以及随机森林文本分类算法在实际案例中的应用。主要实现了垃圾短信的精确识别，通过获得以上挖掘结果，为相关运营商提供一种解决垃圾短信过滤问题的方案。